Enhancing Chat Language Models by Scaling High-quality Instructional Conversations

Year: 2023 | Citations: 721 | Authors: Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng

---

Abstract

Fine-tuning on instruction data has been widely validated as an effective practice for implementing chat language models like ChatGPT. Scaling the diversity and quality of such data, although straightforward, stands a great chance of leading to improved performance. This paper aims to improve the upper bound of open-source models further. We first provide a systematically designed, diverse, informative, large-scale dataset of instructional conversations, UltraChat, which does not involve human queries. Our objective is to capture the breadth of interactions that a human might have with an AI assistant and employs a comprehensive framework to generate multi-turn conversation iteratively. UltraChat contains 1.5 million high-quality multi-turn dialogues and covers a wide range of topics and instructions. Our statistical analysis of UltraChat reveals its superiority in various key metrics, including scale, average length, diversity, coherence, etc., solidifying its position as a leading open-source dataset. Building upon UltraChat, we fine-tune a LLaMA model to create a powerful conversational model, UltraLLaMA. Our evaluations indicate that UltraLLaMA consistently outperforms other open-source models, including Vicuna, the previously recognized state-of-the-art open-source model. The dataset and the model will be publicly released\footnote{\url{https://github.com/thunlp/UltraChat}}.